دیتای مورد بررسی ما مربوط به دیتای بازار آمریکاست. هدف بررسی شاخص NYA در دیتا و تارگت قیمت نهایی (Adj Close) می باشد.

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	12/31/1965	528.690002	528.690002	528.690002	528.690002	528.690002	0.0
1	NYA	1/3/1966	527.210022	527.210022	527.210022	527.210022	527.210022	0.0
2	NYA	1/4/1966	527.840027	527.840027	527.840027	527.840027	527.840027	0.0
3	NYA	1/5/1966	531.119995	531.119995	531.119995	531.119995	531.119995	0.0
4	NYA	1/6/1966	532.070007	532.070007	532.070007	532.070007	532.070007	0.0

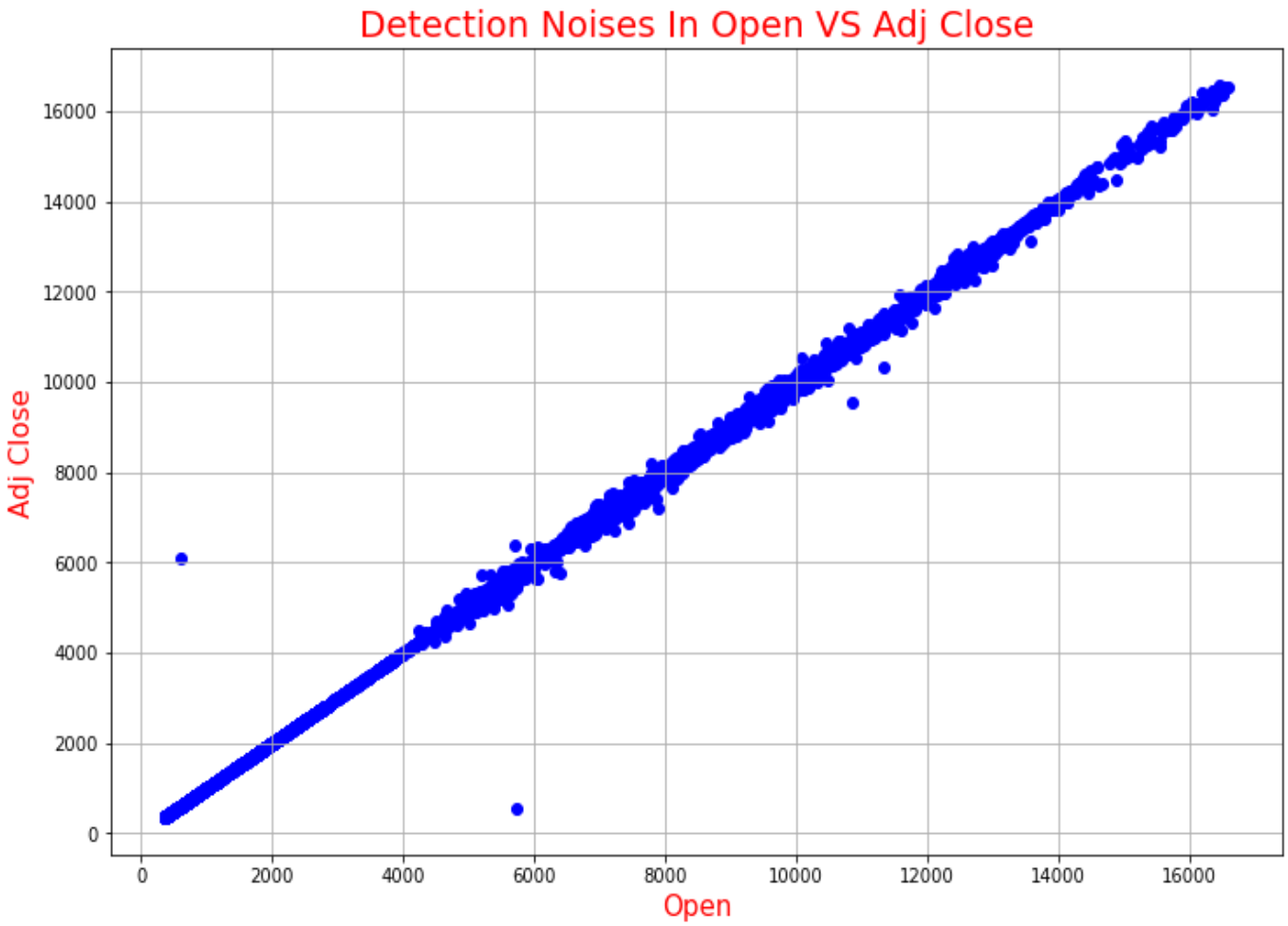
شکل بالا ۵ سطر اول از این دیتاست را نمایش میدهد.

این دیتاست ۱۱۲۴۵۷ سطر و ۸ ستون دارد که اگر شاخص NYA را از آن جدا کنیم ، دیتاستی با ۱۳۹۴۸ سطرو ۸ ستون ایجاد می شود که در شکل زیر مشخص است.

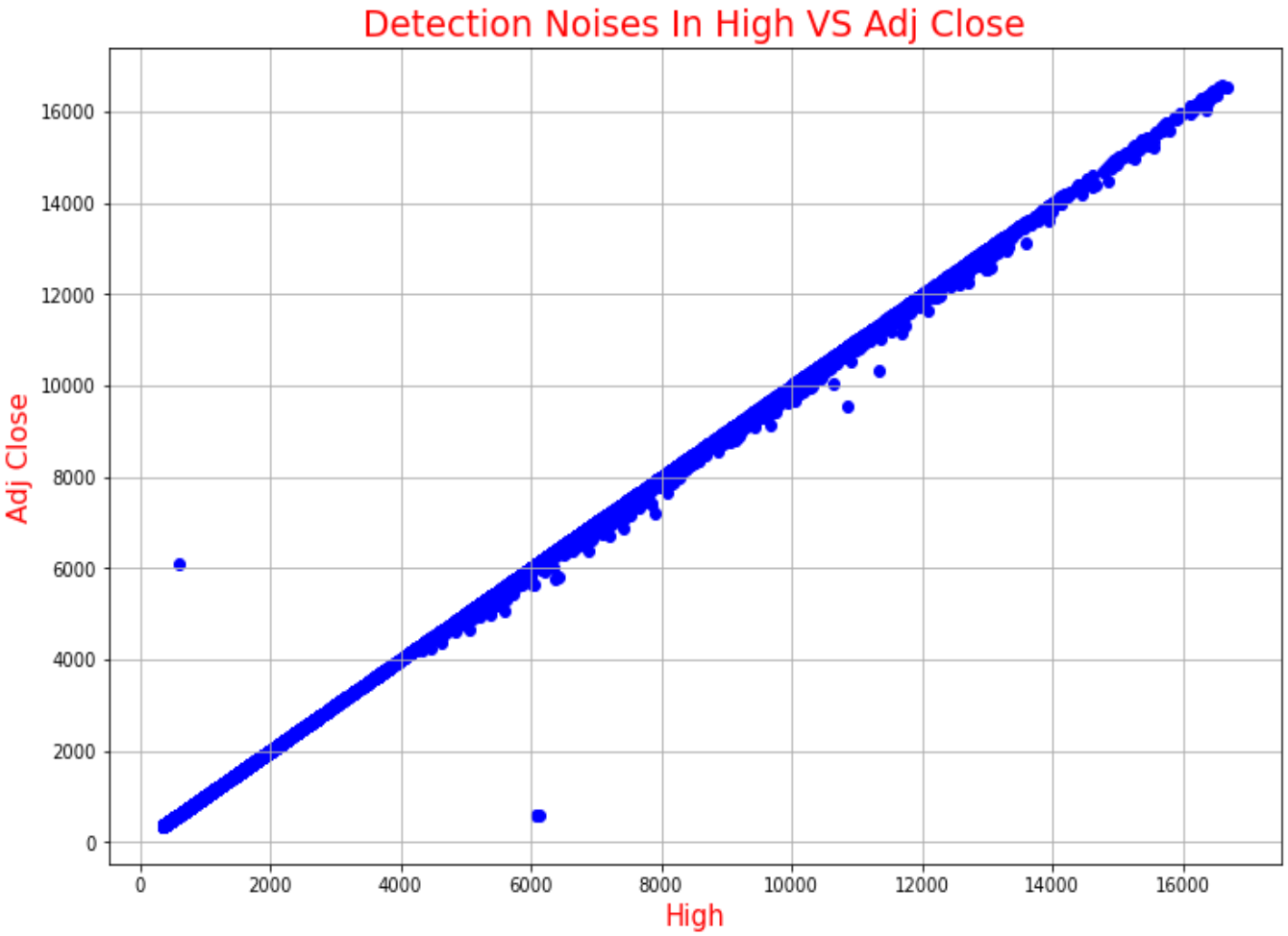
	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	12/31/1965	528.690002	528.690002	528.690002	528.690002	528.690002	0.000000e+00
1	NYA	1/3/1966	527.210022	527.210022	527.210022	527.210022	527.210022	0.000000e+00
2	NYA	1/4/1966	527.840027	527.840027	527.840027	527.840027	527.840027	0.000000e+00
3	NYA	1/5/1966	531.119995	531.119995	531.119995	531.119995	531.119995	0.000000e+00
4	NYA	1/6/1966	532.070007	532.070007	532.070007	532.070007	532.070007	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	5/24/2021	16375.000000	16508.519530	16375.000000	16464.689450	16464.689450	2.947400e+09
13944	NYA	5/25/2021	16464.689450	16525.810550	16375.150390	16390.189450	16390.189450	3.420870e+09
13945	NYA	5/26/2021	16390.189450	16466.339840	16388.320310	16451.960940	16451.960940	3.674490e+09
13946	NYA	5/27/2021	16451.960940	16546.359380	16451.960940	16531.949220	16531.949220	5.201110e+09
13947	NYA	5/28/2021	16531.949220	16588.689450	16531.949220	16555.660160	16555.660160	4.199270e+09

13948 rows × 8 columns

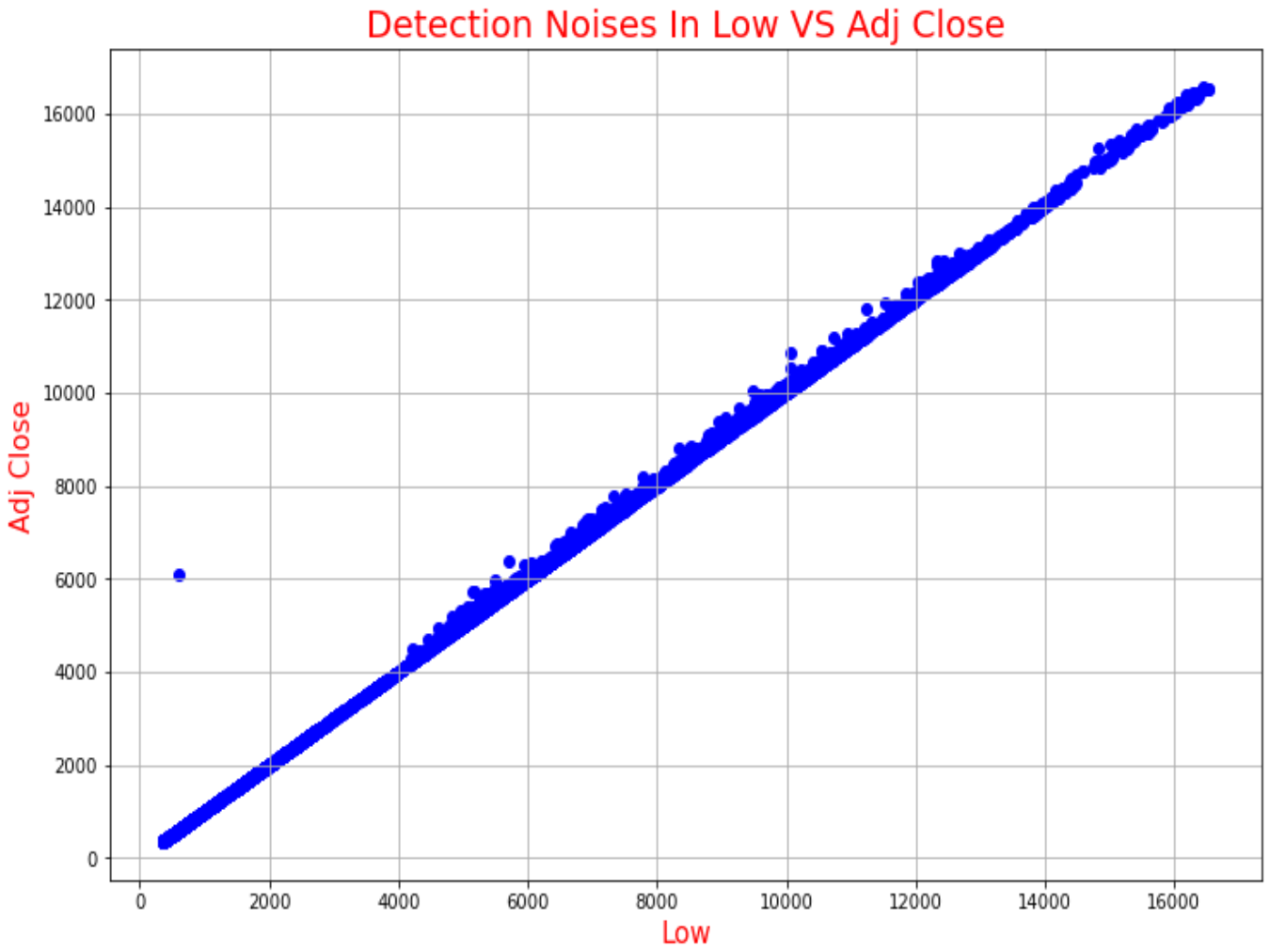
بررسی دو به دوی فیچرها با تارگت برای تشخیص داده های پرت :



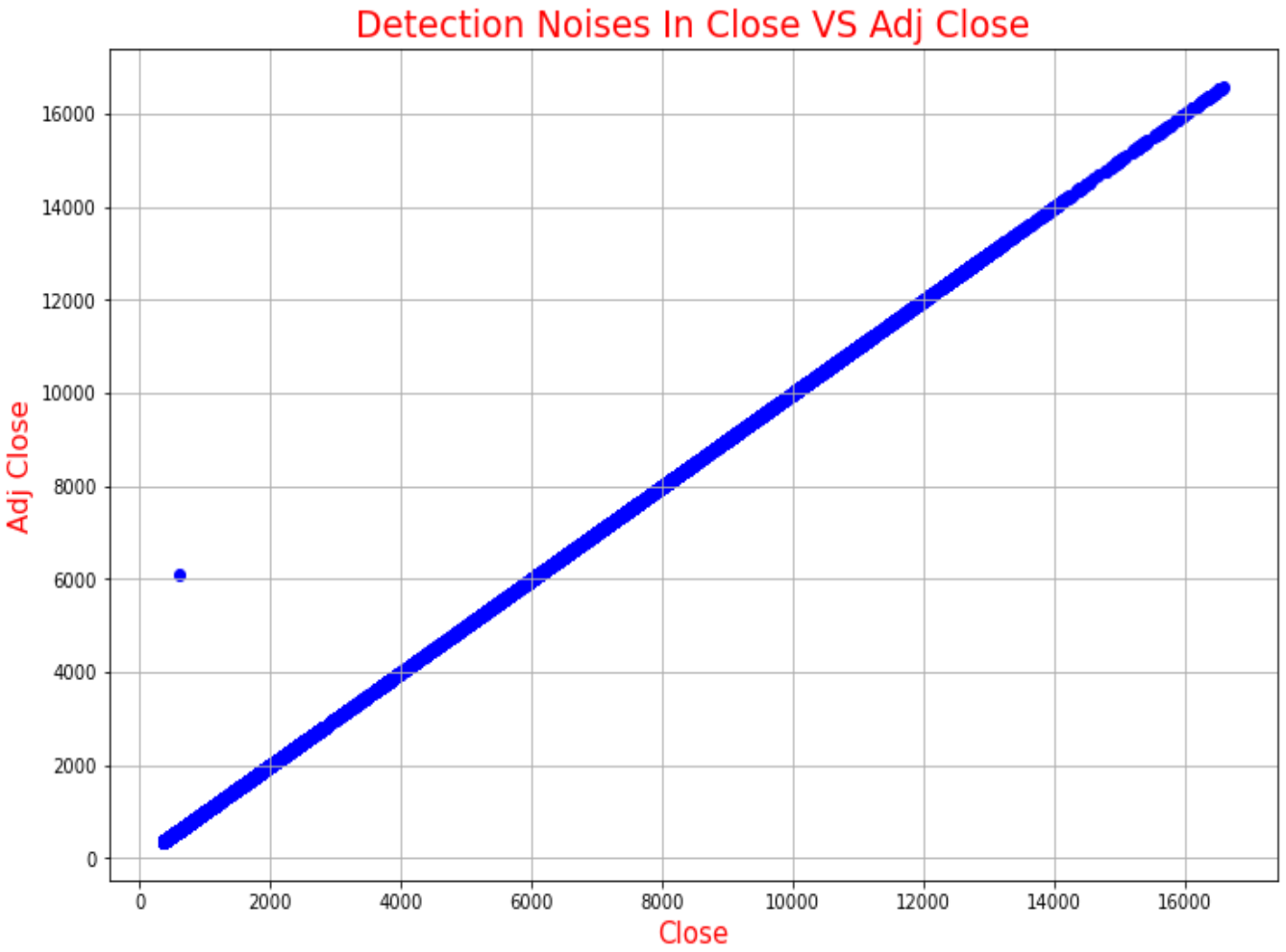
شکل بالا مقایسه قیمت شروع را با قیمت نهایی انجام میدهد ، نمودار بیانگر وجود یک رابطه خطی بین قیمت شروع با قیمت نهایی است همچنین وجود چند داده پرت در نمودار مشخص است که نیاز به بررسی دارند و اگر داده مهمی نبودند یا در اثر اشتباه در اندازه گیری ثبت شده بودند باید حذف شوند.



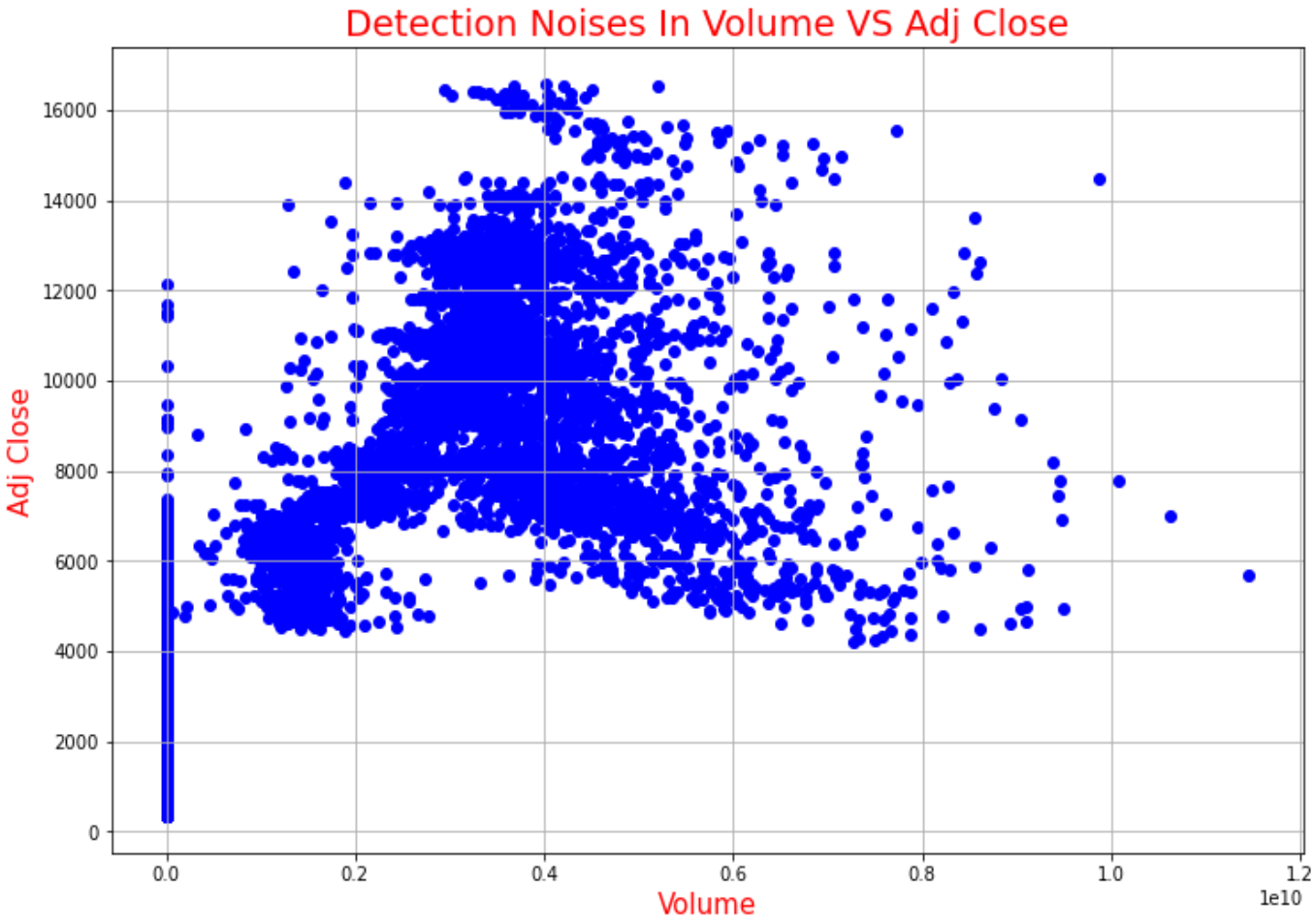
شکل بالا مقایسه بالاترین قیمت را با قیمت نهایی انجام میدهد ، همانطور که از نمودار مشخص است بین بالاترین قیمت و قیمت نهایی یک رابطه تقریبا خطی وجود دارد ، همچنین علاوه بر داده های پرت موجود در پلات قبلی ، وجود چند داده پرت دیگر در قیمت های بیشتر از ۶۰۰۰ و قیمت نهایی کمتر از ۱۰۰۰ مشهود است که نیاز به بررسی دارند که در اینصورت یا داده های بسیار مهمی هستند که باید در تحلیل و مدلسازی اعمال شوند یا در صورت عدم اهمیت باید از دیتاست حذف شوند .



نمودار بالا مقایسه پایین ترین قیمت را با قیمت نهایی انجام میدهد ، نمودار بیانگر وجود یک رابطه خطی بین پایین ترین قیمت با قیمت نهایی است ، همچنین وجود یک داده پرت در نمودار مشخص است که در پلات های قبلی هم وجود داشت که باید پس از بررسی درجه اهمیت آن یا حذف شود یا باقی بماند .



نمودار بالا مقایسه قیمت بسته شدن را با قیمت نهایی انجام میدهد ، نمودار بیانگر وجود یک رابطه کاملا خطی بین قیمت بسته شدن با قیمت نهایی است ، همچنین وجود یک داده پرت در نمودار مشخص است که در پلات های قبلی هم وجود داشت که باید پس از بررسی درجه اهمیت آن یا حذف شود یا باقی بماند .



نمودا بالا مقایسه مقدار والیوم با قیمت نهایی را انجام میدهد ، پس از بررسی همانطور که در نمودار هم مشخص است مشخص شد که ۸۸۳۲ تا از داده ها با مقدار والیوم صفر وجود دارند درصورتیکه مقدار حجم صفر ندارند ، با توجه به اینکه بیش از تقریبا ۵۰ درصد دیتاهای ما مقدار حجمی برابر با صفر دارند ، دو حالت بوجود می آید :

۱- در صورتیکه ما حتما بخواهیم تاثیر این فیچر را در تحلیل و مدلسازی خود ببینیم باید این ۸۸۳۲ سطر با مقدار حجم صفر را از دیتاست حذف کنیم و تحلیل را با باقی داده ها انجام دهیم.

۲- در صورتیکه اثر آن اهمیت چندانی برای ما ندارند کل ستون والیوم را حذف می کنیم.

در این پروژه ستون Volume به طور کلی حذف شده است و دیتاست به شکل زیر درآمده است.

	Index	Date	Open	High	Low	Close	Adj Close
0	NYA	12/31/1965	528.690002	528.690002	528.690002	528.690002	528.690002
1	NYA	1/3/1966	527.210022	527.210022	527.210022	527.210022	527.210022
2	NYA	1/4/1966	527.840027	527.840027	527.840027	527.840027	527.840027
3	NYA	1/5/1966	531.119995	531.119995	531.119995	531.119995	531.119995
4	NYA	1/6/1966	532.070007	532.070007	532.070007	532.070007	532.070007

تشخیص و حذف داده های گمشده :

	Open	High	Low	Close	Adj Close
count	13947.000000	13946.000000	13945.000000	13944.000000	13938.000000
mean	4452.147406	4469.312526	4434.262223	4453.026486	4455.094446
std	4074.835507	4094.956718	4052.815490	4075.483921	4075.456765
min	347.769989	347.769989	347.769989	347.769989	347.769989
25%	654.989990	655.150024	655.039978	655.122513	655.807525
50%	2631.909912	2632.280029	2631.909912	2632.015014	2633.015014
75%	7339.489990	7376.315063	7277.509766	7339.397583	7342.787598
max	16590.429690	16685.890630	16531.949220	16590.429690	16590.429690

همانطور که از سطر count برای ستون های مختلف مشخص است ، تعداد داده های ستون های مختلف با یکدیگر متفاوت است و این به معنای وجود داده های گمشده در بین داده های دیتاست ما است.

```
Show the missing values of DataFrame:

 Index         0
Date           0
Open           1
High           2
Low            3
Close          4
Adj Close     10
dtype: int64
```

خروجی بالا مقدار داده ی گمشده در هر ستون را نمایش می دهد. در مواجه با داده های گمشده چندین رویکرد وجود دارد.

در صورتیکه تعداد داده های گمشده در یک ستون به نسبت کل داده های آن ستون زیاد باشد بررسی می کنیم ، اگر تاثیر آن ستون بر تحلیل و مدلسازی ما برای ما مهم باشد سعی می کنیم داده های گمشده را از روش های مختلف تهیه کرده و جایگزین کنیم در غیر اینصورت کل ستون را حذف می کنیم (مانند ستون Volume) اما اگر تعداد داده های گمشده به نسبت کل داده های آن ستون کم باشد ، داده های گمشده را حذف می کنیم.

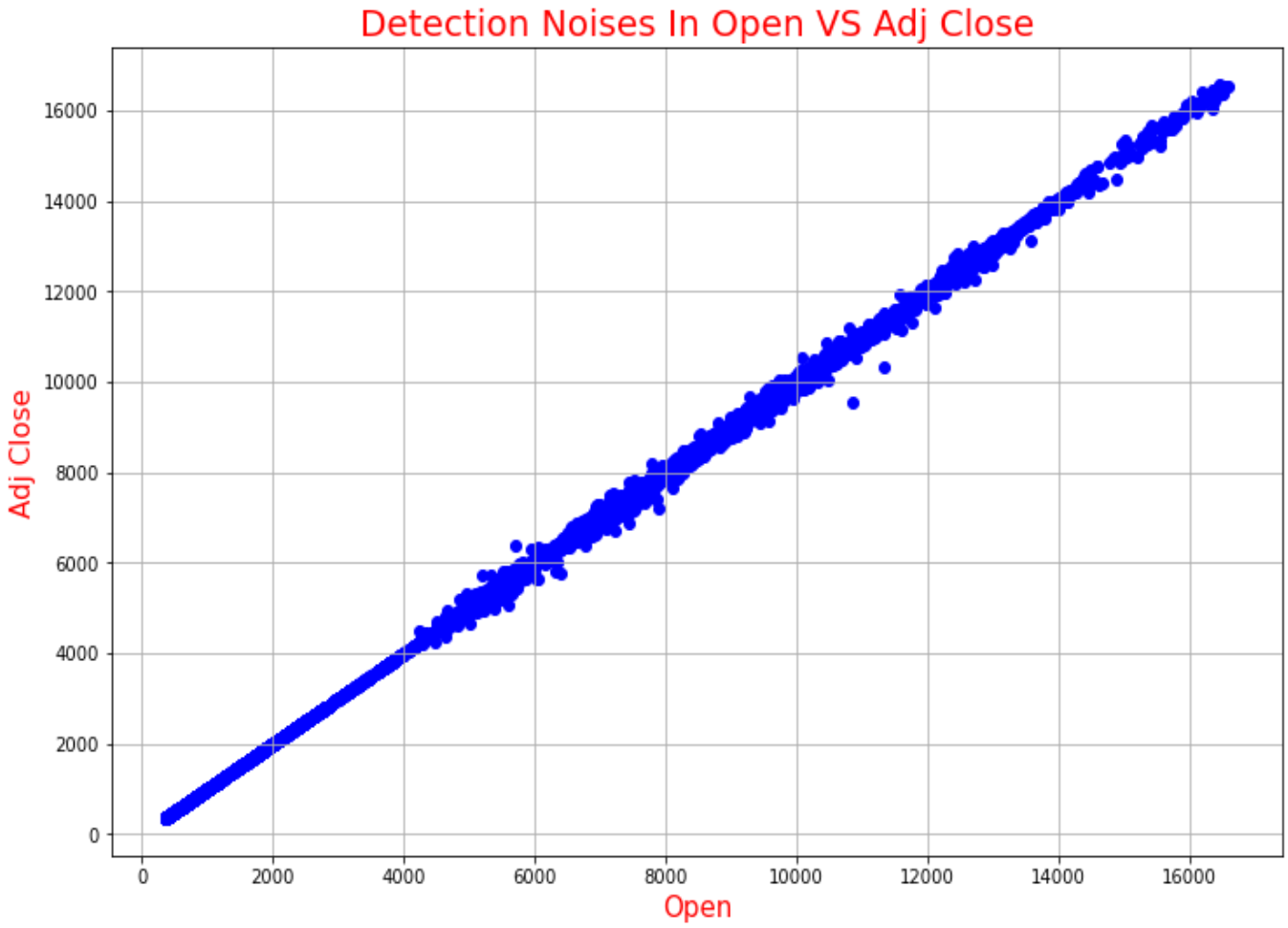
در این دیتاست همانطور که بررسی کردیم تعداد داده های گمشده (۲۰) به نسبت کل داده های دیتاست ما (۱۳۹۴۸) بسیار ناچیز است پس می توانیم آن ها را حذف کنیم.

	Open	High	Low	Close	Adj Close
count	13932.000000	13932.000000	13932.000000	13932.000000	13932.000000
mean	4456.384785	4473.278937	4437.912527	4456.412120	4456.806823
std	4074.980183	4095.100323	4052.942601	4075.604663	4075.498588
min	347.769989	347.769989	347.769989	347.769989	347.769989
25%	656.285004	656.392487	656.182511	656.182511	656.285004
50%	2634.024903	2634.179931	2633.919922	2633.919922	2634.024903
75%	7348.750122	7384.077514	7281.324829	7346.784912	7346.784912
max	16590.429690	16685.890630	16531.949220	16590.429690	16590.429690

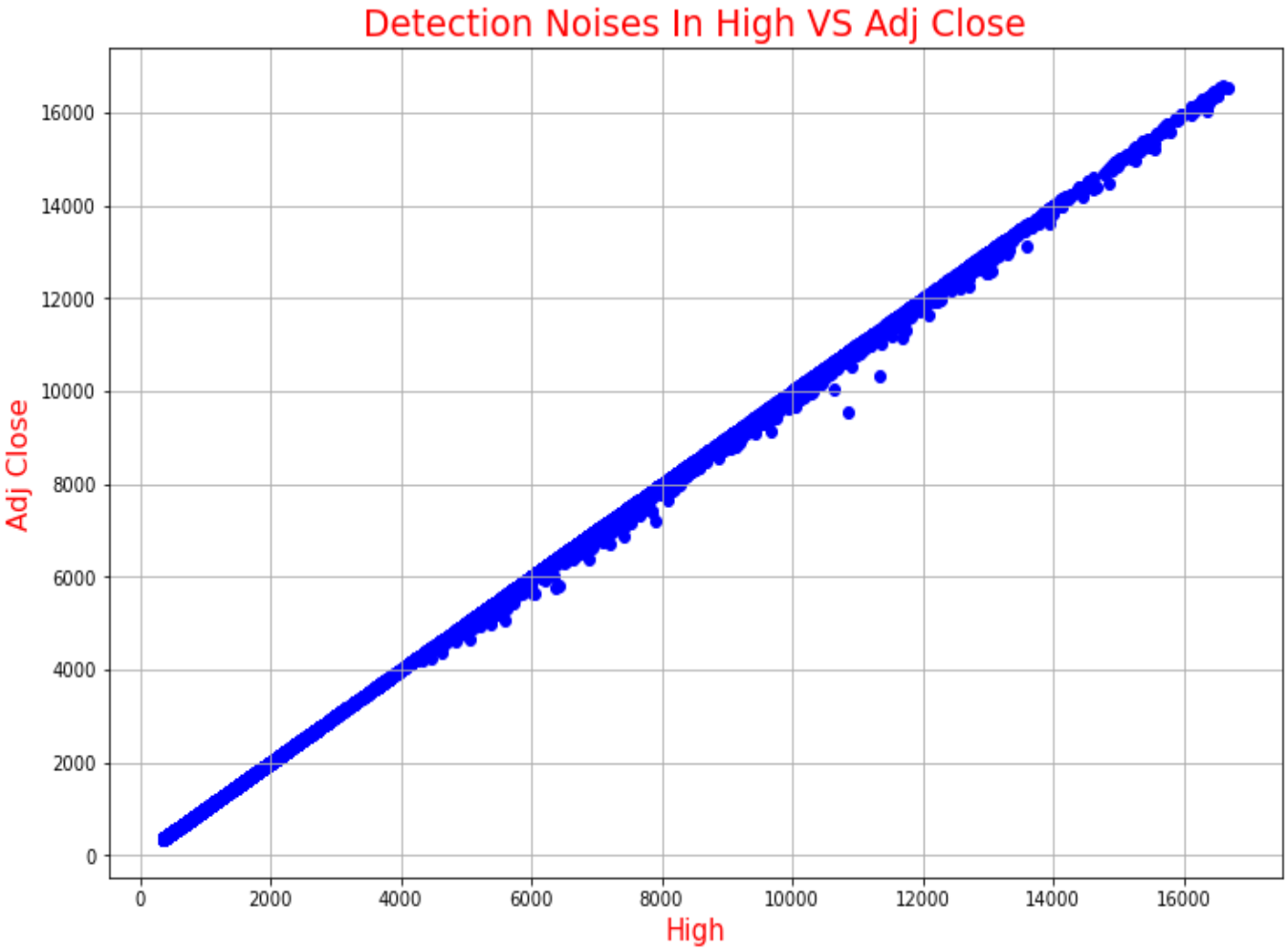
در شکل بالا تغییرات مقدار count ستون های مختلف پس از حذف داده های گمشده نشان داده شده است همانطور که مشخص است پس از حذف داده های گمشده تعداد دیتاها در تمام ستون ها یکسان شده است.

حذف داده های پرت :

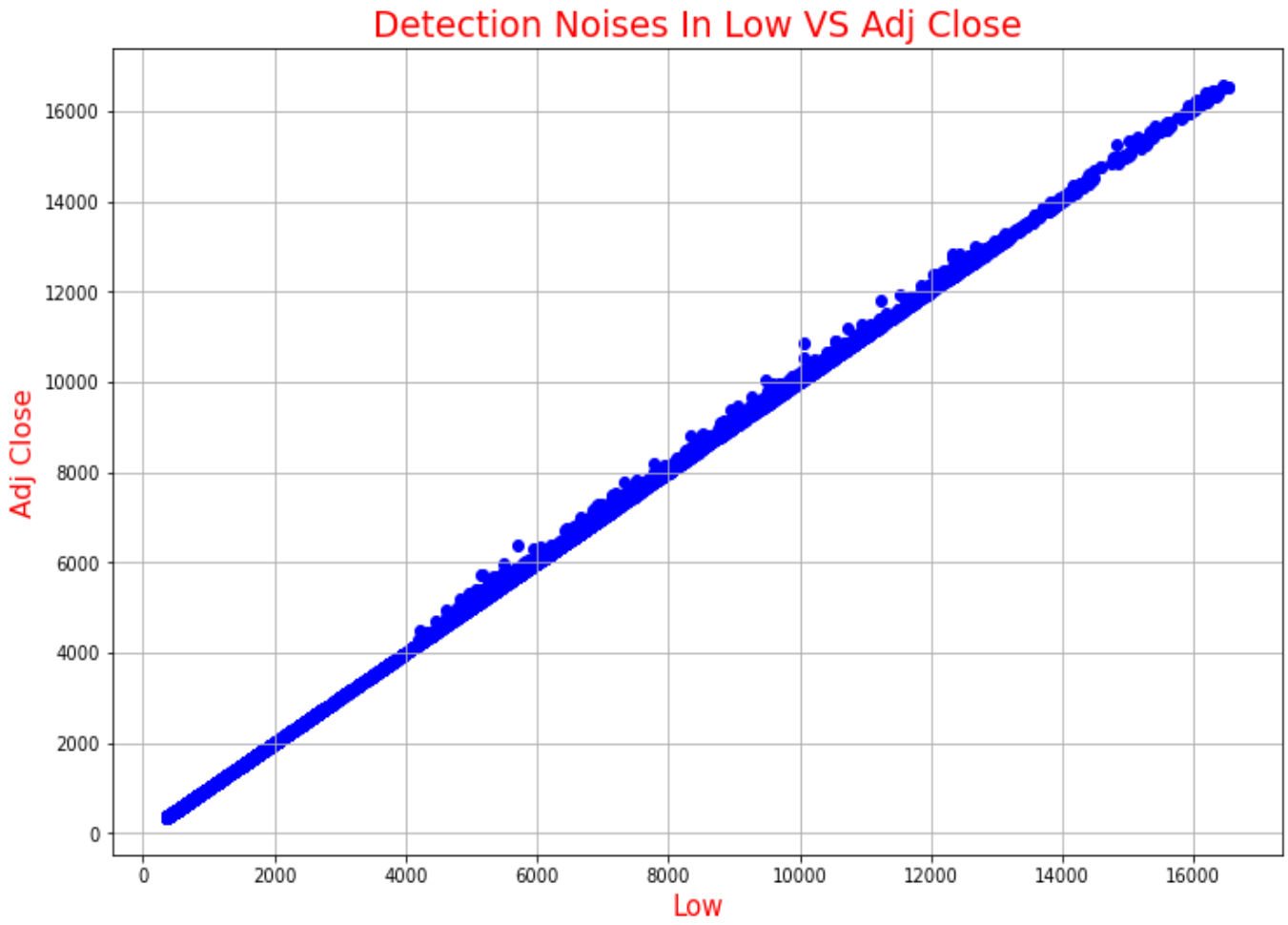
حال نوبت به حذف داده های پرت می رسد ، ابتدا محل داده های پرت را پیدا می کنیم سپس آنها را حذف می کنیم و در نهایت برای بررسی حذف شدن داده های پرت دوباره نمودار ها را ترسیم می کنیم.



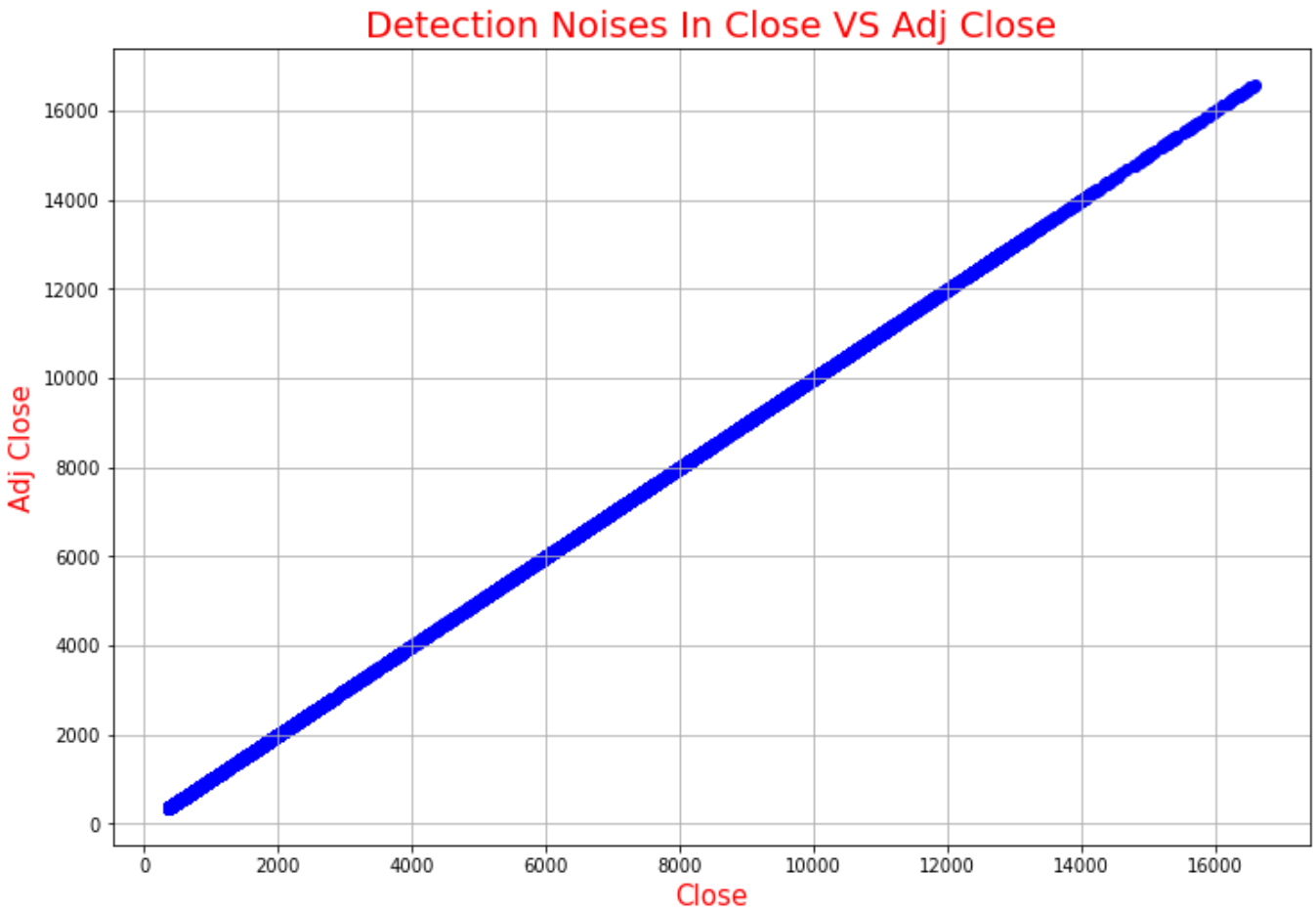
همانطور که مشخص است داده های پرت از این نمودار حذف شده اند.



همانطور که از نمودار مشخص است داده های پرت حذف شده اند.



همانطور که از شکل مشخص است داده های پرت حذف شده اند.



همانطور که از شکل مشخص است داده های پرت حذف شده اند.

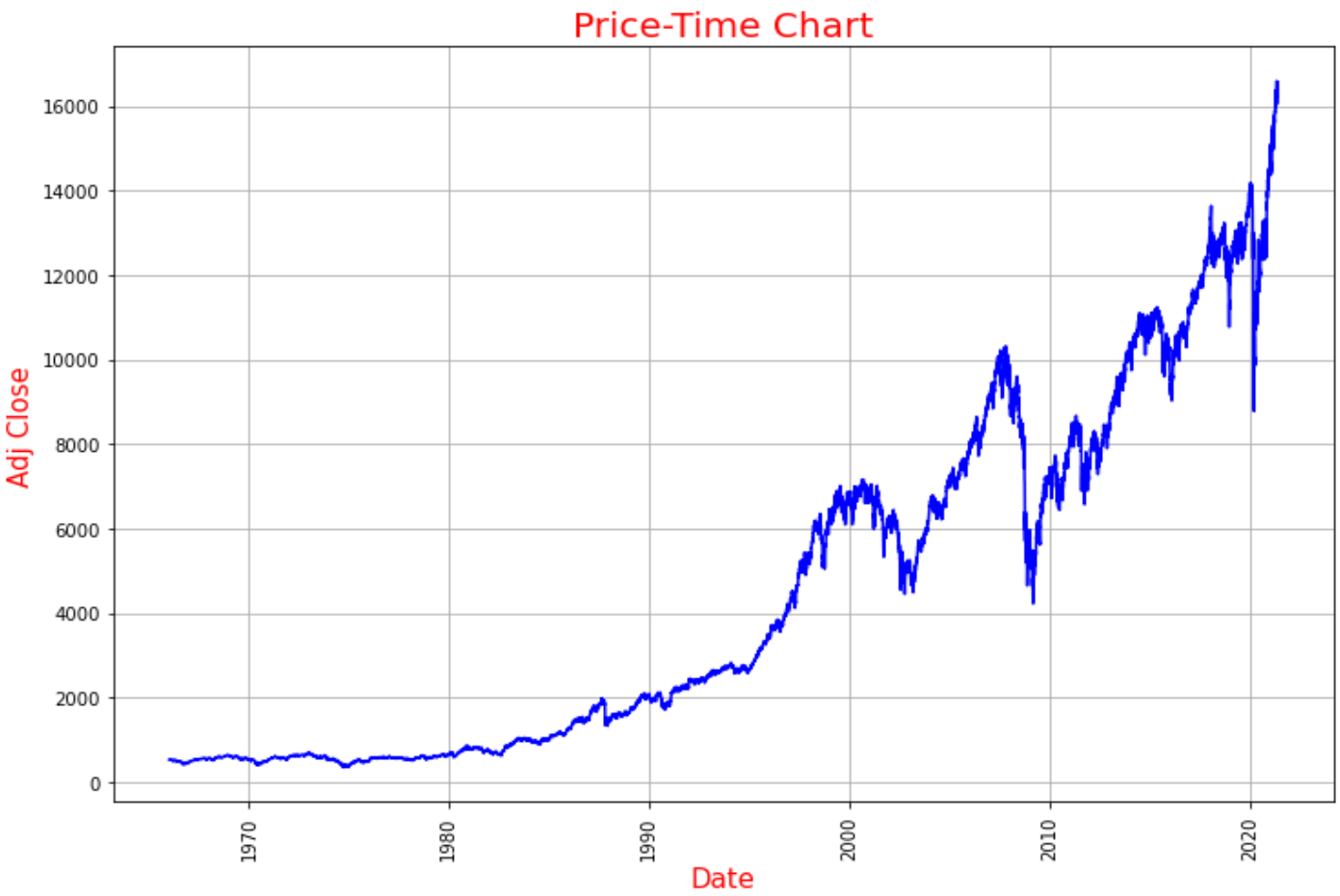
تبدیل اعداد از اعشار به صحیح :

در این مرحله ابتدا دو ستونی که object هستند را دراپ می کنیم سپس بعد از رند کردن اعداد اعشاری آنها را به عدد صحیح تبدیل کرده و درنهایت ۲ ستون حذف شده را به آنها اضافه می کنیم تا دیتاست به شکل زیر تبدیل شود.

	Index	Date	Open	High	Low	Close	Adj Close
0	NYA	12/31/1965	529	529	529	529	529
1	NYA	1/3/1966	527	527	527	527	527
2	NYA	1/4/1966	528	528	528	528	528
3	NYA	1/5/1966	531	531	531	531	531
4	NYA	1/6/1966	532	532	532	532	532
...	...	...	...	...	...	...	...
13923	NYA	5/18/2021	16375	16509	16375	16465	16465
13924	NYA	5/19/2021	16465	16526	16375	16390	16390
13925	NYA	5/20/2021	16390	16466	16388	16452	16452
13926	NYA	5/21/2021	16452	16546	16452	16532	16532
13927	NYA	5/24/2021	16532	16589	16532	16556	16556

13928 rows × 7 columns

آنالیز خطی و نمودار قیمت زمان :



نمودار بالا بیانگر این است که تا سال ۱۹۸۰ بازار آمریکا با یک تعادل و بالانسی در قیمت نهایی کمتر از ۱۰۰۰ فراز و نشیب داشته است از سال ۱۹۸۰ تا ۱۹۹۵ به مدت تقریبی ۱۵ سال قیمت روندی صعودی با شیب ملایم به خود گرفته است اما از سال ۱۹۹۵ به بعد شیب رشد قیمت افزایش بیشتری پیدا میکند و تا سال ۲۰۰۰ به بیشترین مقدار خود میرسد در سال ۲۰۰۳ بنا به دلایل سیاسی یا اقتصادی مختلف قیمت دچار افت شدید می شود اما باز در سال ۲۰۰۷ با شیب بسیار زیاد رو به افزایش می رود تا در سال ۲۰۰۹ بازهم بنا بر دلایل مختلف دچار ریزش شدید شده و به کمترین میزان خود در ۱۰ سال اخیر رسیده است ، از سال ۲۰۱۰ به بعد تقریبا با یک شیب ثابت افزایش قیمت به صورت صعودی بوده است و بازار از لحاظ ریزش و افزایش قیمت خیلی غیرقابل پیش بینی نبوده است غیر از سال ۲۰۲۰ که احتمالا به دلیل همه گیری کرونا دچار ریزش و افت شده است.